# Modelling Dynamic Scenes by Registrating Multi-View Image Sequences

Jean-Philippe Pons, Renaud Keriven, Olivier Faugeras

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Modelling Dynamic Scenes by Registrating Multi-View Image Sequences*

Jean-Philippe Pons — Renaud Keriven — Olivier Faugeras

**N° 5321**

Septembre 2004

Thème COG


ISSN 0249-6399 ISRN INRIA/RR--5321--FR+ENG

# Modelling Dynamic Scenes by Registrating Multi-View Image Sequences

Jean-Philippe Pons , Renaud Keriven , Olivier Faugeras

Thème COG —Systèmes cognitifs
Projet Odyssée

Rapport de recherche n° 5321 —Septembre 2004 — 20 pages

**Abstract:** We present a new variational method for multi-view stereovision and non-rigid three-dimensional motion estimation from multiple video sequences. Our method minimizes the prediction error of the estimated shape and motion. Both problems then translate into a generic image registration task. The latter is entrusted to a similarity measure chosen depending on imaging conditions and scene properties. In particular, our method can be made robust to appearance changes due to non-Lambertian materials and illumination changes. Our method results in a simpler, more flexible, and more efficient implementation than existing deformable surface approaches. The computation time on large datasets does not exceed thirty minutes. Moreover, our method is compliant with a hardware implementation with graphics processor units. Our stereovision algorithm yields very good results on a variety of datasets including specularities and translucency. We have successfully tested our scene flow algorithm on a very challenging multi-view video sequence of a non-rigid event.

**Key-words:** stereovision, non-rigid 3D motion, scene flow, registration, prediction error, variational method, cross correlation, mutual information, non-Lambertian, level sets.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, BP 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : +33 4 92 38 77 77 —Télécopie : +33 4 92 38 77 65

# Modélisation de Scènes Dynamiques par Recalage de Séquences d'Images Multi-Caméras

**Résumé :** Nous présentons une nouvelle méthode variationnelle pour la stéréovision multi-caméras et l'estimation du mouvement tridimensionnel non-rigide à partir de plusieurs séquences vidéos. Notre méthode minimise l'erreur de prédiction de la forme et du mouvement estimés. Les deux problèmes se ramènent alors à une tâche générique de recalage d'images. Cette dernière est confiée à une mesure de similarité choisie en fonction des conditions de prise de vue et des propriétés de la scène. En particulier, notre méthode peut être rendue robuste aux changements d'apparence dus aux matériaux non-lambertiens et aux changements d'illumination. Notre méthode aboutit à une implémentation plus simple, plus souple et plus efficace que les approches par déformation de surface existantes. Le temps de calcul sur de gros jeux de données ne dépasse pas trente minutes. De plus, notre méthode est compatible avec une implémentation matérielle à l'aide de cartes graphiques. Notre algorithme de stéréovision donne de très bons résultats sur de nombreux jeux de données comportant des spécularités et des transparences. Nous avons testé avec succès notre algorithme d'estimation du mouvement sur une séquence vidéo multi-caméras d'une scène non-rigide.

**Mots-clés :** stéréovision, mouvement 3D non-rigide, recalage, erreur de prédiction, méthode variationnelle, corrélation croisée, information mutuelle, non-lambertien, ensembles de niveaux.

# 1 Introduction

Recovering the geometry of a scene from several images taken from different viewpoints, namely *stereovision*, is one of the oldest problems in computer vision. More recently, some authors have considered estimating the dense non-rigid three-dimensional motion field of a scene, often called *scene flow* [28], from multiple video sequences. Both problems require to match different images of the same scene. This is a very difficult task because a scene patch generally has different shapes and appearances when seen from different points of view and over time. To overcome this difficulty, most existing stereovision and scene flow algorithms rely on unrealistic simplifying assumptions that disregard either/both shape/appearance changes.

The brightness constancy assumption is still popular in the stereovision literature, although it requires a precise photometric calibration of the different cameras and only applies to strictly Lambertian scenes. It motivates the multi-view photo-consistency measure used in voxel coloring [23], space carving [13], and in the deformable mesh method of [5]. The variational formulation of [26] proposes an optional local intensity scaling to remove the brightness constancy assumption, but the number of associated partial differential equations varies as the square of the number of cameras, which is often prohibitive.

Similarly, some scene flow methods [30, 3, 15] use the spatio-temporal derivatives of the input images. Due to the underlying brightness constancy assumption and to the local relevance of spatio-temporal derivatives, these differential methods apply exclusively to slowly-moving scenes under constant illumination.

Similarity measures agregating neighborhood information are more robust to noise than point-wise measures and can cope with realistic imaging conditions, but they have to handle geometric distortion between different views and over time. Early methods, like the classical stereovision by correlation, settled for fixed matching windows. The underlying assumption is the fronto parallel hypothesis: camera retinal planes are identical and the scene is an assembly of planes parallel to them. This assumption can still be found in recent work. In [14], the authors disregard projective distortion and attempt to minimize its impact by computing the stereo discrepancy of a scene patch with its two most front-facing cameras only. However, this approach is valid only for a high number of spatially well-distributed cameras.

Some stereovision methods partially handle projective distortion by taking into account the tangent plane to the object [7, 9, 5]. For example, the remarkable approach of [9] allows to estimate both the shape and the non-Lambertian reflectance of a scene by minimizing the rank of a radiance tensor, computed by sampling image intensities on a tesselation of the tangent plane. In such approaches, the matching score depends not only on the position of the surface but also on its orientation. Unfortunately, this first-order shape approximation results in a very complex minimizing flow involving second-order derivatives of the matching score. The computation of these terms is tricky, time-consuming and unstable, and, to our knowledge, all authors have resigned to drop them.

More generally, most techniques trade robustness to realistic photometric conditions for an approximation of shape and motion. This approximation typically behaves like an undesirable regularity constraint and biases the results.

In Section 2, we propose a common variational framework for stereovision and scene flow estimation which correctly handles projective distortion without any approximation of shape and motion and which can be made robust to appearance changes.

The metric used in our framework is the ability to predict the other input views from one input view and the estimated shape or motion. This is related to the methodology proposed in [27] for evaluating the quality of motion estimation and stereo correspondence algorithms. But in our method, the prediction error is used for the estimation itself rather than for evaluation purposes.

Our formulation is completely decoupled from the nature of the image similarity measure used to assess the quality of the prediction. It can be the normalized cross correlation, some statistical measures such as the correlation ratio [19] or the mutual information [29], or any other application-specific measure. Through this choice, we can make the estimation robust to camera spectral sensitivity differences, non-Lambertian materials and illumination changes. In Section 3, we detail two similarity measures that can be used in our framework.

Our method processes entire images from which projective distortion has been removed, thereby avoiding the complex machinery usually needed to match windows of different shapes. Moreover, its minimizing flow is much simpler than in [7, 9]. This results in elegant and efficient algorithms. In Section 4, we describe our implementation. We present our experimental results in Section 5.

## 2 Minimizing the Prediction Error

Our method consists in maximizing, with respect to shape and motion, the similarity between each input view and the predicted images coming from the other views. We adequatly warp the input images to compute the predicted images, which simultaneously removes projective distortion. Numerically, this can be done at a low computational cost using texture-mapping graphics hardware (*cf* Section 4). For example, in the case of stereovision, we reproject the image taken by one camera onto the hypothetical surface, then we predict the appearance of the scene in the other views by projecting this texture-mapped surface in the other cameras. If the estimation of geometry is perfect, the predicted images coincide exactly with the corresponding input images, modulo noise, calibration errors, appearance changes and semi-occluded areas. This motivates our approach: we seek a shape or a motion maximizing the quality of the prediction.

Interestingly, this can be formulated as a generic image registration task. The latter is entrusted to a measure of image similarity chosen depending on imaging conditions and scene properties. This measure is basically a function mapping two images to a scalar value. The more similar the two images are, the lower the value of the measure is.

In our method, we match entire images in the domain of one of the input images. In contrast, some recent works have proposed to estimate the stereo discrepancy by sampling intensities in some neighborhood of the surface, regardless of the resolution of the input images. For example, in [9, 5], the authors resort to a tesselation of the tangent plane, and in [18], the authors perform an integration along the estimated depth map. However, we believe that the similarity measure should be computed in the domain of the input images in order to be faithful to the resolution of the input data. Otherwise, subsampling and interpolation effects may make it irrelevant.

Furthermore, contrarily to [7, 9, 5, 14], our method is not a minimal surface approach, i.e. our energy functional is not written as the integral on the unknown surface of a data fidelity criterion. In this approach, the data attachment term and the regularizing term are mixed whereas we may have to control them separately. As a consequence, to design non trivial regularity constraints, one has to twist the metric. A good discussion of this topic can be found in [25]. The authors show in some numerical experiments that better results can be achieved by integrating the similarity on the images rather than on the surface.

Consequently, in our method, the energy is defined as the sum of a matching term computed in the images and of a regularity constraint. The latter is required to make the problem well-posed. It is application-specific. For example, it could be designed to preserve shape or motion discontinuities. Here we focus on the design of the matching term and we settle for a straightforward regularization for each problem.

The exact minimization of our energy functionals is computationnally unfeasible due to the huge number of unknowns. Indeed, simulated annealing is extremely slow in practice. Graph cuts algorithms yield a global minimum or a strong local minimum, and have proved to perform very well in some formulations of the stereovision problem [11], but they cannot be applied to an arbitrary energy function [12]. Consequently, we must settle for suboptimal strategies, such as gradient descent, that are highly prone to local minima. Our implementation uses a multi-resolution coarse-to-fine strategy to decrease the probability of getting stuck in irrelevant local minima. Basically, it consists in applying the algorithm to a set of smoothed and subsampled images [1].

## 2.1 Stereovision

In the following, let a surface $S \subset \mathbb{R}^3$ model the shape of the scene. We note $I_i : \Omega_i \subset \mathbb{R}^2 \to \mathbb{R}^d$ the image captured by camera $i$. The perspective projection performed by the latter is denoted by $\Pi_i : \mathbb{R}^3 \to \mathbb{R}^2$. Our method takes into account the visibility of the surface points. In the sequel, we will refer to $S_i$ as the part of $S$ visible in image $i$. The reprojection from camera $i$ onto the surface is denoted by $\Pi_{i,S}^{-1} : \Pi_i(S) \to S_i$. With this notation in hand, the reprojection of image $j$ in camera $i$ via the surface writes $I_j \circ \Pi_j \circ \Pi_{i,S}^{-1} : \Pi_i(S_j) \to \mathbb{R}^d$. We note $M$ a generic measure of similarity between two images.

The matching term $\mathcal{M}$ is the sum of the dissimilarity between each input view and the predicted images coming from all the other cameras. Thus, for each oriented pair of cameras $(i, j)$, we compute the similarity between $I_i$ and the reprojection of $I_j$ in camera $i$ via $S$, on the domain where both are defined, i.e. $\Omega_i \cap \Pi_i(S_j)$, in other words after discarding semi-occluded regions:

$$\mathcal{M}(S) = \sum_i \sum_{j \neq i} \mathcal{M}_{ij}(S) \,, \tag{1}$$

$$\mathcal{M}_{ij}(S) = M|_{\Omega_i \cap \Pi_i(S_j)} \left( I_i \,,\; I_j \circ \Pi_j \circ \Pi_{i,S}^{-1} \right) \,. \tag{2}$$

We now compute the variation of the matching term with respect to an infinitesimal vector displacement $\delta S$ of the surface. Figure 1 displays the camera setup and our notations. We neglect the

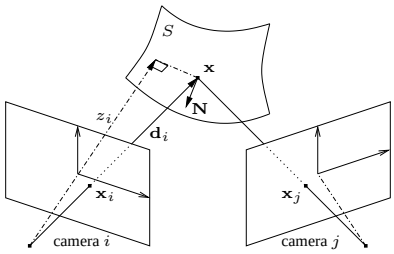


Figure 1: The camera setup and our notations.

variation related to visibility changes. This technical assumption is commonly used in the stereovision literature [7, 9, 5, 14]. Using the chain rule, we get that

$$\left.\frac{\partial \mathcal{M}_{ij}(S+\epsilon\,\delta S)}{\partial \epsilon}\right|_{\epsilon=0} = \int_{\Omega_i \cap \Pi_i(S_j)} \underbrace{\partial_2 M(\mathbf{x}_i)}_{1\times d} \underbrace{DI_j(\mathbf{x}_j)}_{d\times 2} \underbrace{D\Pi_j(\mathbf{x})}_{2\times 3} \underbrace{\left.\frac{\partial \Pi^{-1}_{i,S+\epsilon\,\delta S}(\mathbf{x}_i)}{\partial \epsilon}\right|_{\epsilon=0}}_{3\times 1} d\mathbf{x}_i \ ,$$

where $\mathbf{x}_i$ is the position in image $i$ and $D\cdot$ denotes the Jacobian matrix of a function. For convenience to the reader, we have indicated the dimensions of the different matrices in the product.

When the surface moves, the predicted image changes. Hence the variation of the matching term involves the derivative of the similarity measure with respect to its second argument, denoted by $\partial_2 M$. The meaning of this derivative is detailed in Section 3. Throughout this section, for sake of conciceness, we have omitted the images for which this derivative is evaluated. But the reader must be aware that the predicted images, as well as the domains where the similarity measures are computed, change along the minimizing flow.

We then use a relation between the movement of the surface and the displacement of the reprojected surface point $\mathbf{x} = \Pi^{-1}_{i,S}(\mathbf{x}_i)$:

$$\left.\frac{\partial \Pi^{-1}_{i,S+\epsilon\,\delta S}(\mathbf{x}_i)}{\partial \epsilon}\right|_{\epsilon=0} = \frac{\mathbf{N}^T \delta S(\mathbf{x})}{\mathbf{N}^T \mathbf{d}_i}\,\mathbf{d}_i \ ,$$

where $\mathbf{d}_i$ is the vector joining the center of camera $i$ and $\mathbf{x}$, and $\mathbf{N}$ is the outward surface normal at this point.

Finally, we rewrite the integral in the image as an integral on the surface by the change of variable

$$d\mathbf{x}_i = -\frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3} d\mathbf{x} \ ,$$

where $z_i$ is the depth of $\mathbf{x}$ in camera $i$, and we obtain

$$\left.\frac{\partial \mathcal{M}_{ij}(S+\epsilon\,\delta S)}{\partial \epsilon}\right|_{\epsilon=0} = -\int_{S_i \cap S_j} \left[\partial_2 M(\mathbf{x}_i) DI_j(\mathbf{x}_j) D\Pi_j(\mathbf{x}) \frac{\mathbf{d}_i}{z_i^3}\right] \left[\mathbf{N}^T \delta S(\mathbf{x})\right] d\mathbf{x} \ .$$

In other words, the gradient of the matching term is

$$\nabla \mathcal{M}_{ij}(S)(\mathbf{x}) = -\delta_{S_i \cap S_j}(\mathbf{x}) \left[ \partial_2 M(\mathbf{x}_i) DI_j(\mathbf{x}_j) D\Pi_j(\mathbf{x}) \frac{\mathbf{d}_i}{z_i^3} \right] \mathbf{N} \, , \tag{3}$$

where $\delta_{\cdot}$ is the Kronecker symbol. As expected, the gradient is zero in the regions not visible from both cameras. The reader should also note that the term between square brackets is a scalar function.

As regards the regularization term, it is typically the area of the surface, and the associated minimizing flow is a mean curvature motion. Then, the evolution of the surface is driven by

$$\frac{\partial S}{\partial t} = \left[ -\lambda \, H + \sum_i \sum_{j \neq i} \delta_{S_i \cap S_j} \, \partial_2 M \, DI_j \, D\Pi_j \frac{\mathbf{d}_i}{z_i^3} \right] \mathbf{N} \, , \tag{4}$$

where $H$ denotes the mean curvature of $S$, and $\lambda$ is a positive weighting factor.

## 2.2 Scene flow

Two types of methods prevail in the scene flow literature. In the first family of methods [24, 28, 30], scene flow is constructed from previously computed optical flows in all the input images. However, the latter may be noisy and/or physically inconsistent through cameras. The heuristic spatial smoothness constraints applied to optical flow may also alter the recovered scene flow.

The second family of methods [30, 3, 15] relies on spatio-temporal image derivatives. However, due to the underlying brightness constancy assumption, and to the local relevance of spatio-temporal derivatives, these differential methods apply mainly to slowly-moving Lambertian scenes under constant illumination.

Our method does not fall into these two categories. It directly evolves a 3D vector field to register the input images captured at different times. It can recover large displacements thanks to the multi-resolution strategy and can be made robust to illumination changes through the design of the similarity measure.

Let now $S^t$ model the shape of the scene and $I_i^t$ be the image captured by camera $i$ at time $t$. Let $\mathbf{v}^t : S^t \to \mathbb{R}^3$ be a 3D vector field representing the motion of the scene between $t$ and $t+1$.

The matching term $\mathcal{F}$ is the sum over all cameras of the dissimilarity between the images taken at time $t$ and the corresponding images at $t+1$ warped back in time using the scene flow.

$$\mathcal{F}(\mathbf{v}^t) = \sum_i \mathcal{F}_i(\mathbf{v}^t) \, , \tag{5}$$

$$\mathcal{F}_i(\mathbf{v}^t) = M \left( I_i^t \, , \, I_i^{t+1} \circ \Pi_i \circ (\Pi_{i,S^t}^{-1} + \mathbf{v}^t) \right) \, . \tag{6}$$

As the reader can check easily, its gradient writes

$$\nabla^T \mathcal{F}_i(\mathbf{v}^t)(\mathbf{x}) = -\delta_{S_i^t}(\mathbf{x}) \, \frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3} \, \underbrace{\partial_2 M(\mathbf{x}_i)}_{1 \times d} \, \underbrace{DI_i^{t+1}\left(\Pi_i(\mathbf{x}+\mathbf{v}^t)\right)}_{d \times 2} \, \underbrace{D\Pi_i(\mathbf{x}+\mathbf{v}^t)}_{2 \times 3} \, . \tag{7}$$

As regards the regularization term, in this case it is typically the harmonic energy of the flow over the surface, and the corresponding minimizing flow is an intrinsic heat equation [2]. Then, the evolution of the scene flow is driven by

$$\frac{\partial \mathbf{v}^t}{\partial \tau} = \mu \, \Delta_{S^t} \mathbf{v}^t + \sum_i \delta_{S_i^t} \, \frac{\mathbf{N}^T \mathbf{d}_i}{z_i^3} \, \left[ \partial_2 M \, DI_i^{t+1} \, D\Pi_i \right]^T \ , \tag{8}$$

where $\tau$ is the fictious time of the minimization, $\Delta_{S^t}$ denotes the Laplace-Beltrami operator on the surface, and $\mu$ is a positive weighting factor.

# 3 Some Similarity Measures

In this section, we present two similarity measures than can be used in our framework: cross correlation and mutual information [29]. Cross correlation assumes a local affine dependency between the intensities of the two images, whereas mutual information can cope with general statistical dependencies. We have picked these two measures among a broader family of statistical criteria proposed in [8] for multimodal image registration.

In the following, we consider two scalar images $I_1, \, I_2 : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$. The measures below can be extended to vector (e.g. color) images by summing over the different components.

The minimizing flows given in Section 2 involve the derivative of the similarity measure with respect to the second image, denoted by $\partial_2 M$. The meaning of this derivative is the following: given two images $I_1, \, I_2 : \Omega \to \mathbb{R}^d$, we note $\partial_2 M(I_1, I_2)$ the function mapping $\Omega$ to the row vectors of $\mathbb{R}^d$, verifying for any image variation $\delta I$:

$$\left. \frac{\partial M(I_1, I_2 + \epsilon \, \delta I)}{\partial \epsilon} \right|_{\epsilon = 0} = \int_\Omega \partial_2 M(I_1, I_2)(\mathbf{x}) \, \delta I(\mathbf{x}) \, d\mathbf{x} \ . \tag{9}$$

## 3.1 Cross correlation

Cross correlation is still the most popular matching measure in the stereovision area. Most methods still use fixed square or rectangular matching windows. In this case, the choice of the window size is a difficult trade-off between match reliability and oversmoothing of depth discontinuities due to projective distortion [21]. Some authors alleviate this problem by using adaptative windows [10, 20].

In our method, since we match distortion-free images, the size of the matching window is not related to a shape approximation. The matter here is in how big a neighborhood the assumption of affine dependency is valid. Typically, non-Lambertian scenes require to reduce the size of the correlation window, making the estimation less robust to noise and outliers.

In our implementation, we use smooth Gaussian windows with an infinite support instead of hard windows. Gaussian windows are more elegant as regards the continuous formulation of our problem and can be implemented efficiently with fast recursive filtering.

Thus, we gather neighborhood information using convolutions by a Gaussian kernel of standard deviation $\sigma$. The local mean, variance, covariance and cross correlation of the two images respec-

tively write

$$\mu_i(\mathbf{x}) = \frac{G_\sigma \star I_i(\mathbf{x})}{\omega(\mathbf{x})} \ , \qquad v_i(\mathbf{x}) = \frac{G_\sigma \star I_i^2(\mathbf{x})}{\omega(\mathbf{x})} - \mu_i^2(\mathbf{x}) + \beta^2 \ ,$$
$$v_{1,2}(\mathbf{x}) = \frac{G_\sigma \star I_1 I_2(\mathbf{x})}{\omega(\mathbf{x})} - \mu_1(\mathbf{x})\,\mu_2(\mathbf{x}) \ , \quad cc(\mathbf{x}) = \frac{v_{1,2}(\mathbf{x})}{\sqrt{v_1(\mathbf{x}) v_2(\mathbf{x})}} \ ,$$

where $\omega$ is a normalization function accounting for the shape of the domain: $\omega(\mathbf{x}) = \int_\Omega G_\sigma(\mathbf{x} - \mathbf{y})\, d\mathbf{y}$. The $\beta$ constant prevents the denominator from being zero.

We agregate the opposite of the local cross correlation to get a similarity measure corresponding to our needs:

$$M^{CC}(I_1, I_2) = -\int_\Omega cc(\mathbf{x})\, d\mathbf{x} \ . \tag{10}$$

The minimizing flow involved by our method includes the derivative of the similarity measure with respect to the second image. In this case, it writes

$$\partial_2 M^{CC}(I_1, I_2)(\mathbf{x}) = \alpha(\mathbf{x})\, I_1(\mathbf{x}) + \beta(\mathbf{x})\, I_2(\mathbf{x}) + \gamma(\mathbf{x}) \ , \tag{11}$$

where

$$\alpha(\mathbf{x}) = G_\sigma \star \frac{-1}{\omega\,\sqrt{v_1\, v_2}}(\mathbf{x}) \ \ , \ \ \beta(\mathbf{x}) = G_\sigma \star \frac{cc}{\omega\, v_2}(\mathbf{x}) \ \ , \ \ \gamma(\mathbf{x}) = G_\sigma \star \left( \frac{\mu_1}{\omega\,\sqrt{v_1\, v_2}} - \frac{\mu_2\, cc}{\omega\, v_2} \right)(\mathbf{x}) \ .$$

In practice, along the minimizing flow, the $\alpha, \beta, \gamma$ functions change slowly relative to $I_1$ and $I_2$. So, in our implementation, we update them only every ten iterations to reduce the computational burden.

## 3.2 Mutual information

Mutual information is based on the joint probability distribution of the two images, estimated by the Parzen window method [17] with a Gaussian kernel of standard deviation $\beta$:

$$P(i_1, i_2) = \frac{1}{|\Omega|} \int_\Omega G_\beta \left( I_1(\mathbf{x}) - i_1 \ , \ I_2(\mathbf{x}) - i_2 \right) d\mathbf{x} \ . \tag{12}$$

We note $P_1, P_2$ the marginals:

$$P_1(i_1) = \int_{\mathbb{R}} P(i_1, i_2)\, di_2 \quad , \quad P_2(i_2) = \int_{\mathbb{R}} P(i_1, i_2)\, di_1 \ .$$

Our measure is the opposite of the mutual information of the two images:

$$M^{MI}(I_1, I_2) = -\int_{\mathbb{R}^2} P(i_1, i_2)\, \log \frac{P(i_1, i_2)}{P_1(i_1) P_2(i_2)}\, di_1\, di_2 \ . \tag{13}$$

Its derivative with respect to the second image writes [8, 6]:

$$\partial_2 M^{MI}(I_1, I_2)(\mathbf{x}) = \zeta(I_1(\mathbf{x}), I_2(\mathbf{x})) \ , \tag{14}$$

where

$$\zeta(i_1, i_2) = \frac{1}{|\Omega|}\, G_\beta \star \left( \frac{\partial_2 P}{P} - \frac{P_2'}{P_2} \right)(i_1, i_2)\ .$$

In our implementation, the $\zeta$ function is updated only every ten iterations.

# 4 Implementation Aspects

We have implemented our method in the level set framework [4, 16], motivated by its numerical stability and its ability to handle topological changes automatically. However, our method is not specific to a particular surface model. Thus, an implementation with meshes would be straightforward.

The predicted images can be computed very efficiently thanks to graphics card hardware-accelerated rasterizing capabilities. In our implementation, we determine the visibility of surface points in all cameras using OpenGL depth buffering, we compute the reprojection of an image to another camera via the surface using projective texture mapping, and we discard semi-occluded areas using shadow-mapping [22].

The bottleneck in our current implementation is the computation of the similarity measure. Since it only involves homogeneous operations on entire images, we could probably resort to a graphics processor unit based implementation with vertex and fragment programs.

# 5 Experimental Results

## 5.1 Stereovision

Table 1 describes the stereovision datasets used in our experiments. All datasets are color images except “Hervé” which is grayscale. All are real images except “Buddha”. “Cactus” and “Gargoyle” are courtesy of Pr. K. Kutulakos (University of Toronto). “Buddha” and “Bust” are publicly available from the OpenLF software (LFM project, Intel).

We have used either cross correlation (CC) or mutual information (MI). Both perform well on these complex scenes. “Buddha” and “Bust” are probably the more challenging datasets: “Buddha” is a synthetic scene simulating a translucent material and “Bust” includes strong specularities. However, cross correlation with a small matching window (standard deviation of 2 pixels) yields very good results.

Using all possible camera pairs is quite expensive computationnally. Moreover, it is often not necessary since, when two cameras are far apart, no or little part of the scene is visible in both views. Consequently, in practice, we only pick pairs of neighboring cameras.

The computation time is reasonable: up to 30 minutes on a 2 GHz Pentium IV PC under linux. The number of iterations is 600 for all datasets. However, in practice, the convergence is often attained earlier. Hence the computation time could be reduced using an appropriate stopping criterion. In all our experiments, the regularizer is a mean curvature motion.

We show our results in Figures 2,3,4,5 and 6. For each dataset, we display some of the input images, the ground truth when available, then some views of the estimated shape, and finally the

| Name | #Images | Image size | #Image pairs | Measure | Level set size | Time (sec.) |
|---|---|---|---|---|---|---|
| Hervé | 2 | $512 \times 512$ | 2 | MI | $128^3$ | 107 |
| Cactus | 30 | $768 \times 484$ | 60 | CC | $128^3$ | 1670 |
| Gargoyle | 16 | $719 \times 485$ | 32 | MI | $128^3$ | 905 |
| Buddha | 25 | $500 \times 500$ | 50 | CC | $128^3$ | 530 |
| Bust | 24 | $300 \times 600$ | 48 | CC | $128 \times 128 \times 256$ | 1831 |

Table 1: Description of the stereovision datasets used in our experiments.

same views after reprojecting the texture coming from the most front-facing camera. Note that this texture-mapped representation does not aim at photorealism. In particular, it generates artefacts at the places where the source of the texture changes. It is only intended to show the validity of the output of our method for more sophisticated image-based rendering techniques.

In all our experiments, the overall shape of the objects is successfully recovered, and a lot of details are captured: the eyes and the mouth of "Hervé", the stings of "Cactus", the ears and the pedestal of "Gargoyle", the nose and the collar of "Buddha", the ears and the moustache of "Bust". A few defects are of course visible. Some of them can be explained. The hole around the stick of "Gargoyle" is not fully recovered. This may be due to the limited number of images (16): some parts of the concavity are visible only in one camera. The depression in the forehead of "Bust" is related to a very strong specularity: intensity is almost saturated in some images.

Figure 2: "Hervé" stereo pair and our results.

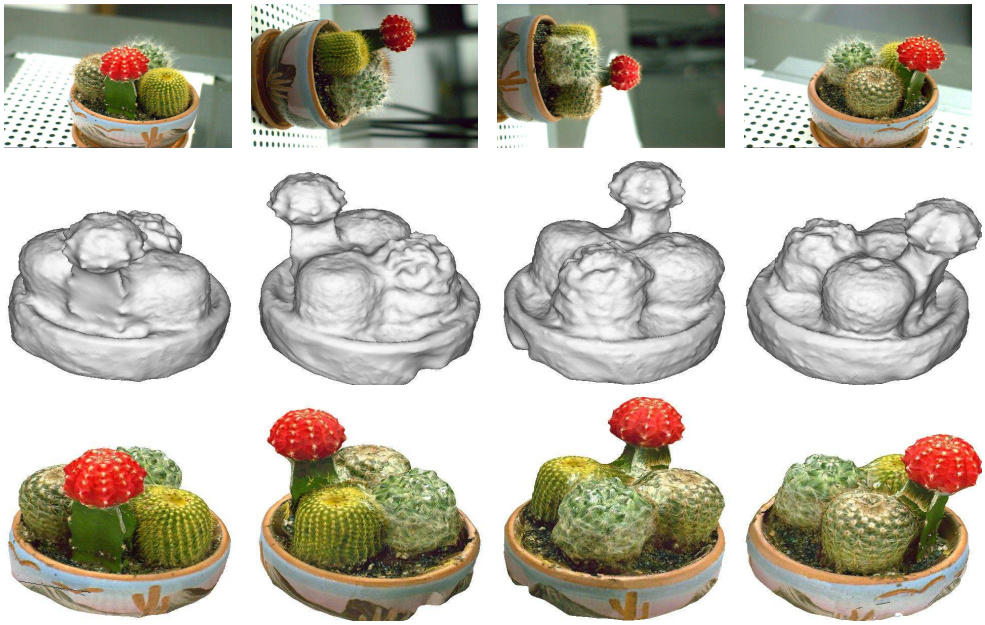

Figure 3: Some images from the “Cactus” dataset and our results.

## 5.2 Stereovision + scene flow

We have tested our scene flow algorithm on a challenging multi-view video sequence of a non-rigid event. The “Yiannis” sequence is taken from a collection of datasets that were made available to the community by P. Baker and J. Neumann (University of Maryland) for benchmark purposes. This sequence shows a character talking while rotating his head. It was captured by 22 cameras at 54 fps plus 8 high-resolution cameras at 6 fps. Here we focus on the 30 synchronized sequences at the lower frame rate to demonstrate that our method can handle large displacements.

We have applied successively our stereovision and scene flow algorithms: once we know the shape $S^t$, we compute the 3D motion $\mathbf{v}^t$ with our scene flow algorithm. Since $S^t + \mathbf{v}^t$ is a very good estimate of $S^{t+1}$, we use it as the initial condition in our stereovision algorithm and we perform a handful of iterations to refine it. This is mush faster than restarting the optimization from scratch. We also compute the backward motion from $t + 1$ to $t$ for the purpose of time interpolation.

Figure 7 displays the first four frames of one of the input sequence and our estimation of shape and 3D forward motion at corresponding times. We successfully recover the opening then closing of the mouth, followed by the rotation of the head while the mouth opens again. Moreover, we capture displacements of more than twenty pixels.

We use our results to generate time-interpolated 3D sequences of the scene. To synthetize images at intermediate time instants, we can either use the previous shape and texture warped by the forward

Figure 4: Some images from the “Gargoyle” dataset and our results.

motion, or the next shape and texture warped by the backward motion. Ideally the two should coincide exactly, but of course this is never the case in practice. As a consequence, we linearly interpolate between forward and backward extrapolated images to guarantee a smooth blending between frames. In return it causes “crossfading” artefacts in some places where forward and backward extrapolation significantly diverge.

We display a short excerpt of such a time-interpolated sequence in Figure 8. Note the progressive opening and closing of the mouth.

# 6 Conclusion

We have presented a novel method for multi-view stereovision and scene flow estimation which minimizes the prediction error. Our method correctly handles projective distortion without any approximation of shape and motion, and can be made robust to appearance changes. To achieve this, we adequatly warp the input views and we register the resulting distortion-free images with a user-defined similarity measure. We have implemented our stereovision method in the level set framework and we have obtained results comparing favorably with state-of-the-art methods, even on complex non-Lambertian real-world images including specularities and translucency. Using our

Figure 5: Some images from the “Buddha” dataset, ground truth and our results.

algorithm for motion estimation, we have successfully recovered the 3D motion of a non-rigid event and we have synthetized time-interpolated 3D sequences.

## Acknowledgements

We would like to thank Pr. Kyros Kutulakos for providing us the “Cactus” and “Gargoyle” datasets, and Dr. Jan Neumann for his support on the “Yiannis” dataset.

Figure 6: Some images from the “Bust” dataset, pseudo ground truth and our results.

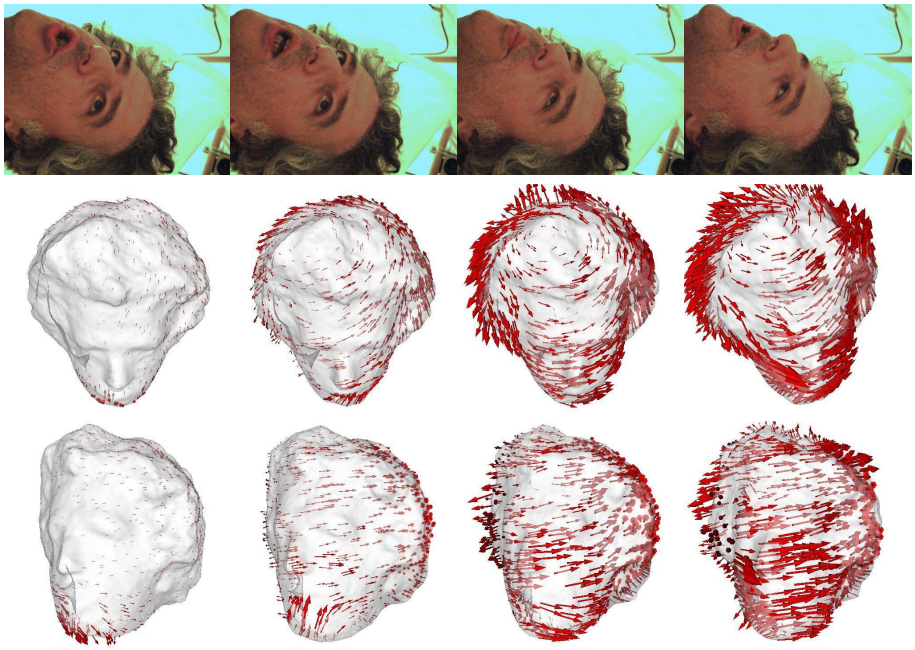

Figure 7: First images of one sequence of the “Yiannis” dataset and our results.

Figure 8: An excerpt of the time-interpolated 3D sequence for the “Yiannis” dataset.

# References

[1] L. Alvarez, J. Weickert, and J. Sànchez. Reliable estimation of dense optical flow fields with large displacements. *The International Journal of Computer Vision*, 39(1):41–56, August 2000.

[2] M. Bertalmio, L.T. Cheng, S. Osher, and G. Sapiro. Variational problems and partial differential equations on implicit surfaces. *Journal of Computational Physics*, 174:759–780, December 2001.

[3] R.L. Carceroni and K.N. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape and reflectance. *The International Journal of Computer Vision*, 49:175–214, 2002.

[4] A. Dervieux and F. Thomasset. A finite element method for the simulation of Rayleigh-Taylor instability. *Lecture Notes in Mathematics*, 771:145–159, 1979.

[5] Y. Duan, L. Yang, H. Qin, and D. Samaras. Shape reconstruction from 3D and 2D data using PDE-based deformable surfaces. In T. Pajdla and J. Matas, editors, *Proceedings of the 8th European Conference on Computer Vision*, volume 3023, pages 238–251, Prague, Czech Republic, 2004. Springer–Verlag.

[6] O. Faugeras and G. Hermosillo. Well-posedness of two non-rigid multimodal image registration methods. *Siam Journal of Applied Mathematics*, 64(5):1550–1587, 2004.

[7] Olivier Faugeras and Renaud Keriven. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):336–344, March 1998.

[8] Gerardo Hermosillo, Christophe Chefd'hotel, and Olivier Faugeras. Variational methods for multimodal image matching. *The International Journal of Computer Vision*, 50(3):329–343, November 2002.

[9] H. Jin, S. Soatto, and A.J. Yezzi. Multi-view stereo beyond Lambert. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 171–178, Madison, Wisconsin (United States), June 2003.

[10] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932, September 1994.

[11] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 7th European Conference on Computer Vision*, volume 3, Copenhagen, Denmark, May 2002. Springer–Verlag.

[12] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Proceedings of the 7th European Conference on Computer Vision*, volume 3, Copenhagen, Denmark, May 2002. Springer–Verlag.

[13] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *The International Journal of Computer Vision*, 38(3):199–218, July 2000.

[14] M. Lhuillier and L. Quan. Surface reconstruction by integrating 3D and 2D data of multiple views. In *Proceedings of the 9th International Conference on Computer Vision*, Nice, France, 2003. IEEE Computer Society, IEEE Computer Society Press.

[15] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *The International Journal of Computer Vision*, 47:181–193, 2002.

[16] S. Osher and J. Sethian. Fronts propagating with curvature dependent speed: algorithms based on the Hamilton–Jacobi formulation. *Journal of Computational Physics*, 79:12–49, 1988.

[17] E. Parzen. On the estimation of probability density function. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[18] J.-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In *Proceedings of the 9th International Conference on Computer Vision*, pages 597–602, Nice, France, 2003. IEEE Computer Society, IEEE Computer Society Press.

[19] Alexis Roche, Grégoire Malandain, Xavier Pennec, and Nicholas Ayache. Multimodal image registration by maximization of the correlation ratio. Technical Report 3378, INRIA, August 1998.

[20] D. Scharstein and R. Szeliski. Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174, June 1998.

[21] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *The International Journal of Computer Vision*, 1(47):7–42, 2002.

[22] M. Segal, C. Korobkin, R. van Widenfelt, J. Foran, and P. Haeberli. Fast shadows and lighting effects using texture mapping. *Computer Graphics*, 26(2):249–252, 1992.

[23] S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. *The International Journal of Computer Vision*, 35(2):151–173, 1999.

[24] Y.Q. Shi, C.Q. Shu, and Pan J.N. Unified optical flow field approach to motion analysis from a sequence of stereo images. *Pattern Recognition*, 27(12):1577–1590, 1994.

[25] S. Soatto, A.J. Yezzi, and H. Jin. Tales of shape and radiance in multi-view stereo. In *Proceedings of the 9th International Conference on Computer Vision*, volume 2, pages 974–981, Nice, France, 2003. IEEE Computer Society, IEEE Computer Society Press.

[26] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide-baseline views. In *Proceedings of the 9th International Conference on Computer Vision*, volume 2, pages 1194–1201, Nice, France, 2003. IEEE Computer Society, IEEE Computer Society Press.

[27] R. Szeliski. Prediction error as a quality metric for motion and stereo. In *Proceedings of the 7th International Conference on Computer Vision*, volume 2, pages 781–788, Kerkyra, Greece, 1999. IEEE Computer Society, IEEE Computer Society Press.

[28] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *Proceedings of the 7th International Conference on Computer Vision*, volume 2, pages 722–729, Kerkyra, Greece, 1999. IEEE Computer Society, IEEE Computer Society Press.

[29] Paul Viola and William M. Wells III. Alignement by maximization of mutual information. *The International Journal of Computer Vision*, 24(2):137–154, 1997.

[30] Y. Zhang and C. Kambhamettu. On 3D scene flow and structure estimation. In *Proceedings of CVPR'01*, 2001.

# Contents

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifi que
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399